Distinguishing Commercial from Editorial Content in News

Timo Kats[1] Peter van der Putten[1] Jasper Schelling[2]

[1]LIACS (Leiden University)

[2]ACED

November 10, 2021

Introduction

- Differentiating commercial and editorial content in the form of ***articles*** and ***advertorials***.
- An ***advertorial*** is commercial content that's formatted as an article, often branded by a small disclaimer.
- This interferes with the ***separation of church and state*** in journalism.
- Research in the context of ***Reverb Channel*** platform, collaboration between Leiden University and ACED, on role reversal for AI in digital media

Separation of church and state in the news

Refers to the clear distinction there should be between the news' editorial and commercial content.

Research questions

- To what extent can we differentiate advertorials and articles by using machine learning?
- Can we use AI and machine learning to better understand the difference between commercial and editorial language?

Explorative research

We also explored the collected data and results to get a better insight through making visualizations.

Data acquisition (overview)

- For every newspaper we (roughly) have 250 advertorials and 250 articles.
- Most articles and advertorials are from 2020 (small bias).
- Difficult to collect more, since advertorials tend to get deleted.

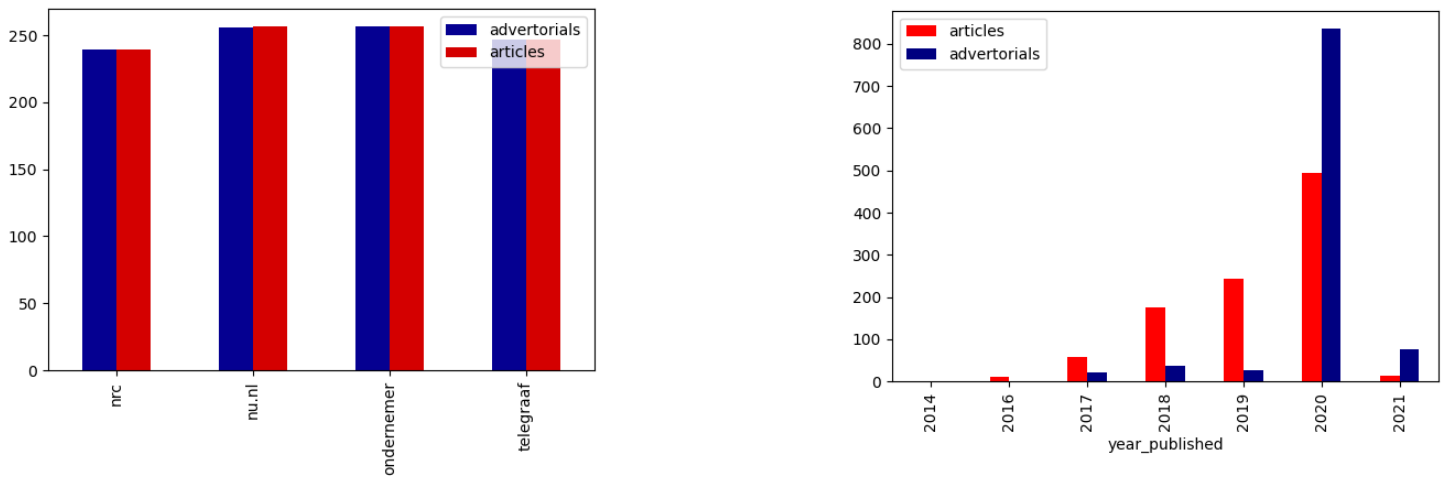


Figure: Metadata from the acquired data set

Data acquisition (scraping)

```
import requests

index = 0
source = ""

article_file = open('nu/articles.txt', 'a+')
ads_file = open('nu/ads.txt', 'a+')

while True:
    url = 'https://www.nu.nl/algemeen/' + str(index)
    request = requests.get(url)
    print("currently at index: ", index)
    if request.status_code == 200:
        source = request.text
        if("Dit is een gesponsord artikel op NU.nl" in source):
            ads_file.write(str(index) + '\n')
            ads_file.flush()
        else:
            article_file.write(str(index) + '\n')
            article_file.flush()
    index += 1
```

Python code for url scraping (nu.nl)

Experimental set up

- Data selection
  - Input: Text of the article/advertorial
  - Target: Is this entry sponsored?
- Data cleaning
  - Removing leakers (brandnames and disclaimers)
  - Converting to lowercase

Results (model)

After experimenting with different models, text representations, number of features and SVM parameters (see paper) we obtained the following results:

Performance of the best performing model:

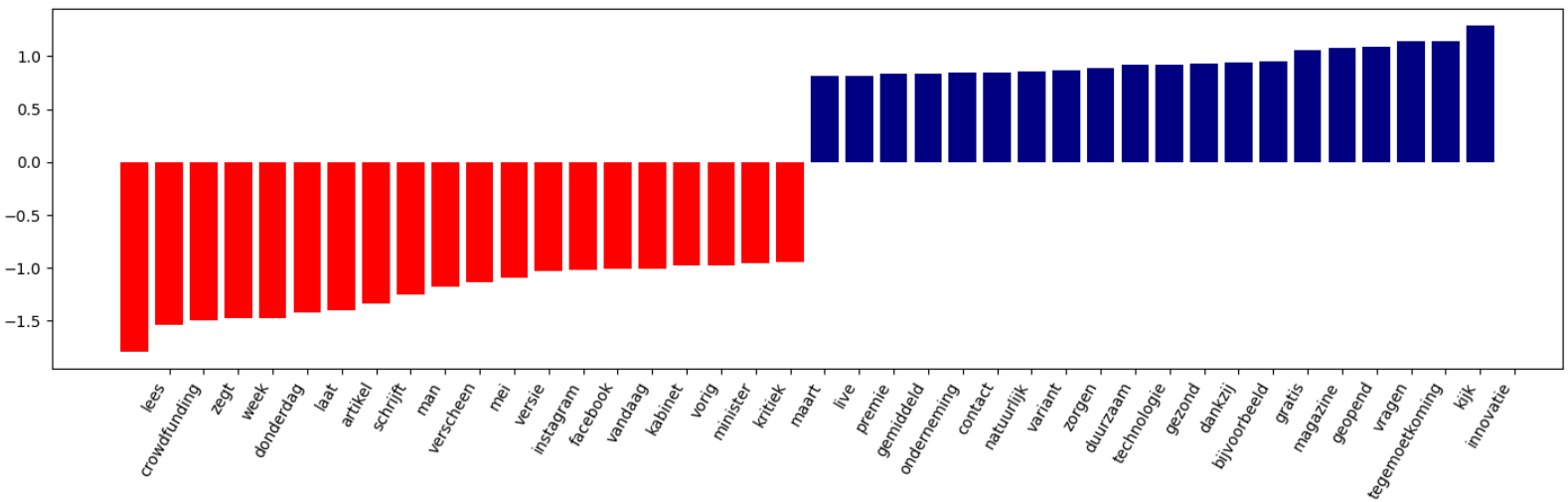
- Accuracy=0.9029±0.0559
- f1-score=0.9003±0.0581
- roc_auc=0.9495±0.0341

Overview of the best performing model:

learning model	**features**	**text representation**	**kernel**	**max amount of iterations**
svm	5000	tf-idf	linear	5000

Results (lexicon)

- Why? Shows the story behind the model (easier to use outside tech).
- What? A dictionary with the 5000 features with scores.
- How? Derived all the (5000) features from our model.



Exploring our data: Co-occurrence graphs

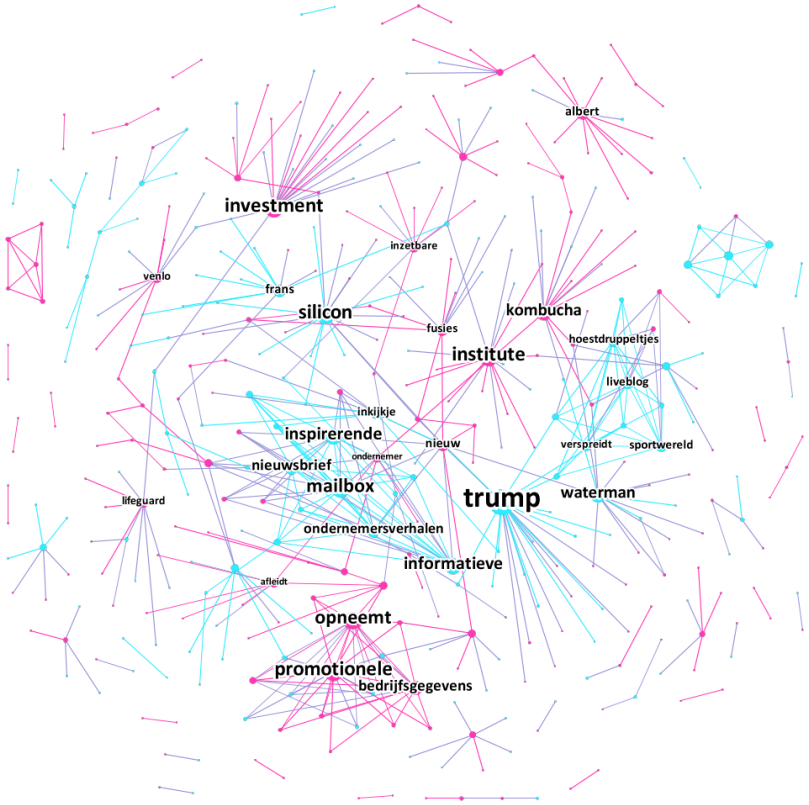


Figure: https://timokats.github.io/network/

Discussion

Back to the separation of church and state...

- Our results can be seen as an indicator.
- Visualized with T-sne. Showing the predicted entries in 2D.
- "NRC" displays clearest separation, "Ondernemer" the least.

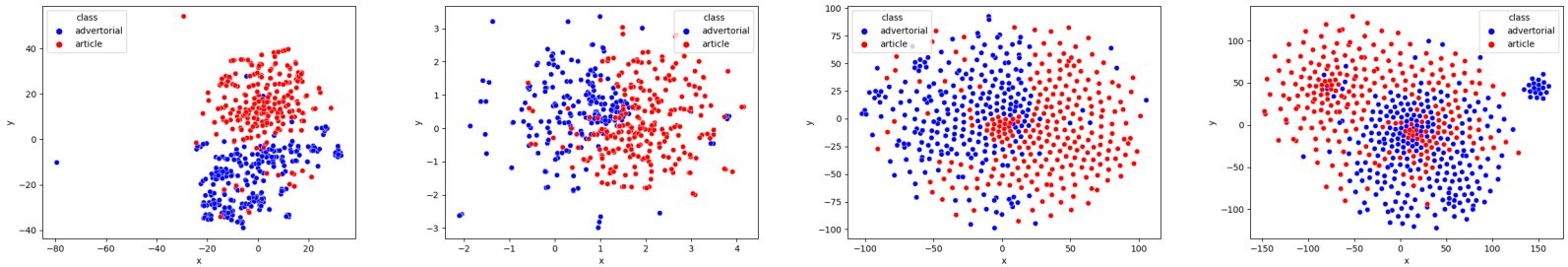


Figure: NRC (95%), Nu.nl (92%), Telegraaf (91%) and Ondernemer (85%)

Further research and recommendations

- Adding a stemmer
- Increase the size of the data set
- Sentiment analysis
- The effect of disclosing advertorials on readers
- Research the effect of different disclaimers with machine learning

Conclusions

- To what extent can we differentiate advertorials and articles by using machine learning? **Best model scored just over 90%.**
- Can we use AI and machine learning to better understand the difference between commercial and editorial language? **Yes.**

However...

- Small data set.

Nevertheless...

- Showed potential.
- Showed positive role AI and ML can fulfill in the modern media landscape.

Thank you